

# Implementing a Data Publishing Service via DSpace

---

Jon W. Dunn, Randall Floyd, Garrett Montanez, Kurt Seiffert

# Outline

- IUScholarWorks
- Massive Data Storage Service
- Example of the data publishing need
- What is the data publishing service
- Conceptual overview of DSpace implementation

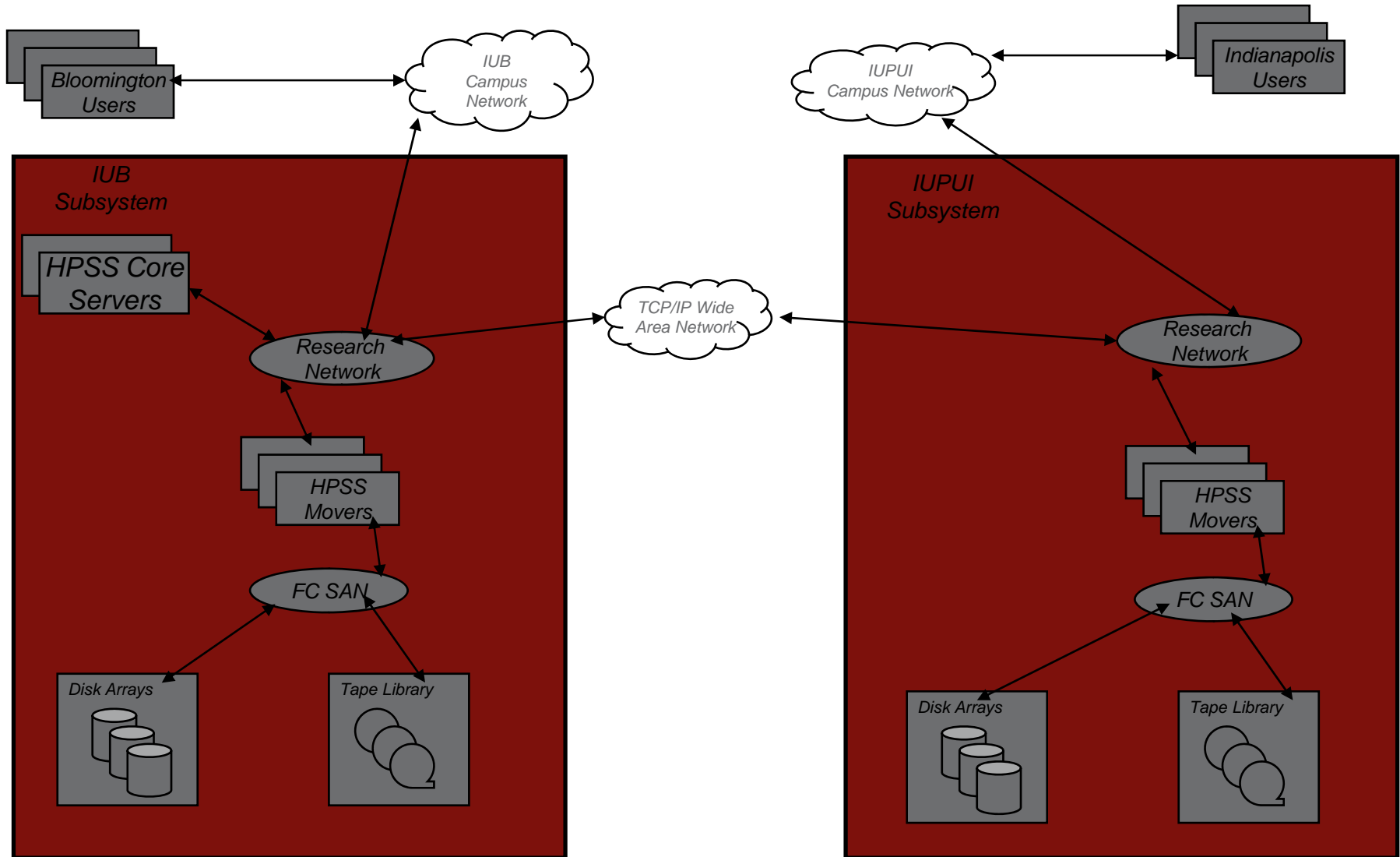
# IUScholarWorks

- IUScholarWorks – Indiana University's (IU's) scholarly communication services
- IUScholarWorks Team – members from IU Libraries and the Digital Library Program
- Current services:
  - A DSpace-based IR - articles, papers, technical reports, etc
  - An Open Journal System-based scholarly journal hosting service

# Overview of MDSS

- Massive Data Storage System (MDSS)
- Current system for research data storage
- Installed in 1998
- Based on IBM developed High Performance Storage System (HPSS) software
- It offers over 2.8 petabytes of disk- and tape-based storage. Distributed between Indianapolis and Bloomington campuses

# ***Distributed between IUB and IUPUI***



# Transferring Files in MDSS

- Fastest Methods
  - hsi
  - Gridftp
  - pftp\_client
  - kerberized ftp
- Convenient Methods
  - Sftp
  - https
  - Samba
  - Hpssfs

# Example of Data Publishing Need

- Linked Environments for Atmospheric Discovery (LEAD)
  - Weather forecasting experiments
  - Want to capture the entire workflow from an experiment
  - Each workflow ~10GB
  - They are looking for a mechanism to preserve the workflows and make them available to others

## Model Domain Configuration

## Region Type Selection

- ☒ Regional 1000Km X 1000Km X 51 Domain with 5 Km Grid Spacing
- ☐ Regional 800Km X 800Km X 51 Domain with 4 Km Grid Spacing
- ☐ Regional 400Km X 400Km X 51 Domain with 2 Km Grid Spacing

## Forecast Start Time

Dates and times in Greenwich Mean Time (GMT)

- ☒ Now (in other words, run a forecast using the most recent data available)
- ☐ Please specify:

Start Date: 2009/05/17 Current Time: 2009/05/17 15:42Z

Start Hour: 0Z

Forecast Duration: 6 hours

Using your mouse, drag and drop the center of the model domain grid to position it as desired on the map





# IUScholarWorks Data

- A new service of the IUScholarWorks repository
- Allow for the publishing of datasets
- Data will have a persistent URL so it can be linked to publications
- The service will combine our DSpace repository with IU's Massive Data Storage system (MDSS), a system that researchers are already uses
- If a file is over a certain size, it will be stored in MDSS
- Allows discovery over the Web
- Preservation – bit level

# Collaborative effort

- IU Libraries
- Research Technologies division - IU's central IT organization, University Information Technology Services (UITS)
- Digital Library Program (a collaboration between the Libraries and UITS)
- IU's Office of the Vice-Provost for Research

# Current Activities

- Two phased implementation
  - Phase one – more manual on the part of the DSpace administrator, user
  - Phase two- more automated system
- Convene focus groups
- Metadata requirements
- DSpace/MDSS integration

# Two scenarios

- Researcher already uses MDSS to store their data
- Researcher does not use MDSS to store their data

# Classes of Files

1. **Small Data Files** – would go directly into DSpace in the underlying asset store as bitstreams
2. **Large Data File**
  1. Preexisting datasets in MDSS account directory
  2. User needs to upload new datasets to MDSS

# Conceptual overview of DSpace implementation

---

# IUScholarWorks Data in DSpace

- Recap of the primary goals of the service:
  - Discovery and access of datasets and related publications through the IUScholarWorks Repository service
  - Facilitating the submission process for both the researcher and collection manager

# IUScholarWorks Data in DSpace

- Discovery and access of datasets and related publications through the IUScholarWorks Repository service
  - DSpace records that are searchable, indexed, and harvested and available at stable URL's
  - DSpace records that contain DSpace bitstreams for small datasets
  - DSpace records that link to large datasets in IU MDSS



# IUScholarWorks Data: Linking to MDSS and delivery via HTTP



Item record with  
URL's of datasets  
in MDSS

HTTP  
Server

hpssfs  
filesystem

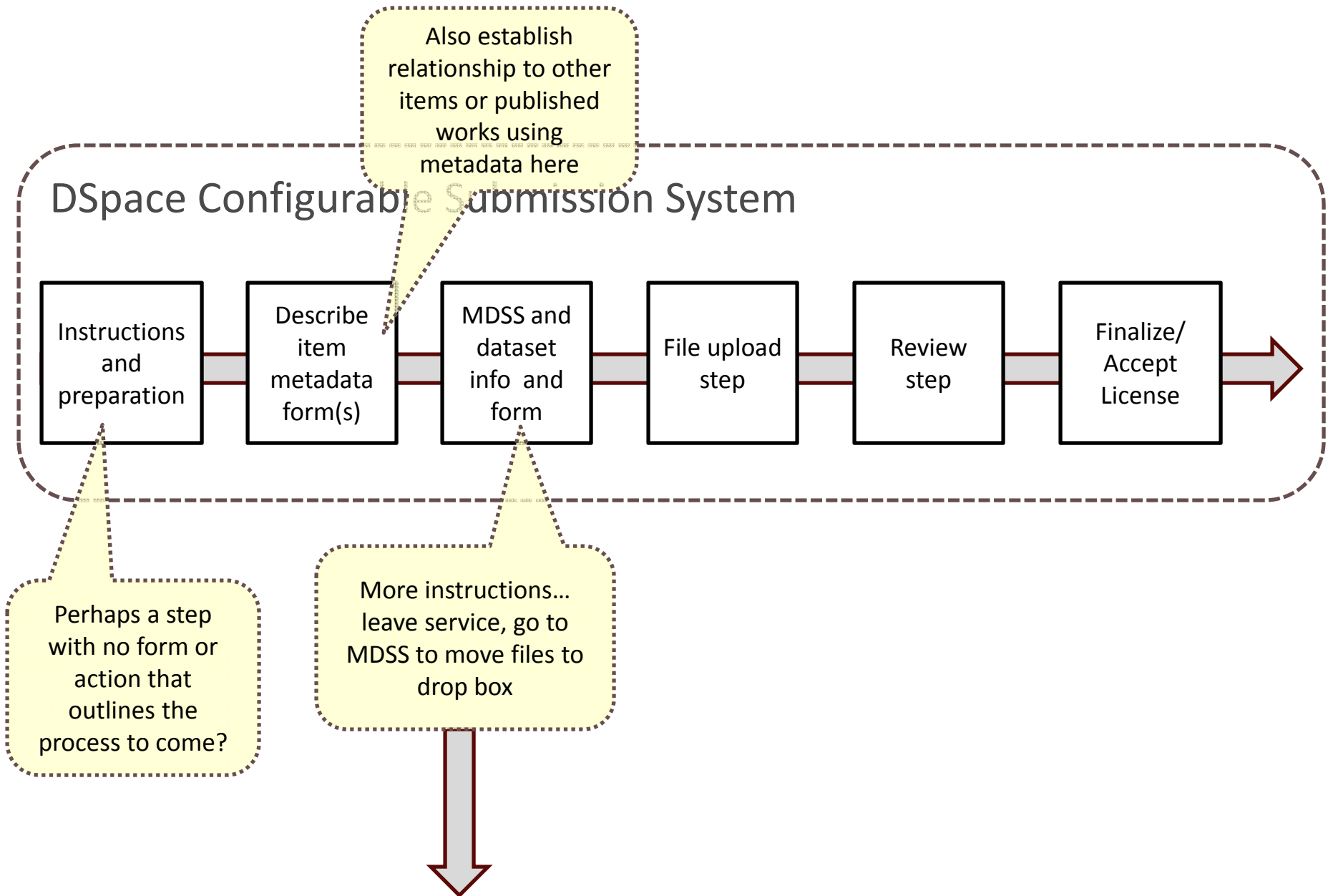
IU MDSS

MDSS web server

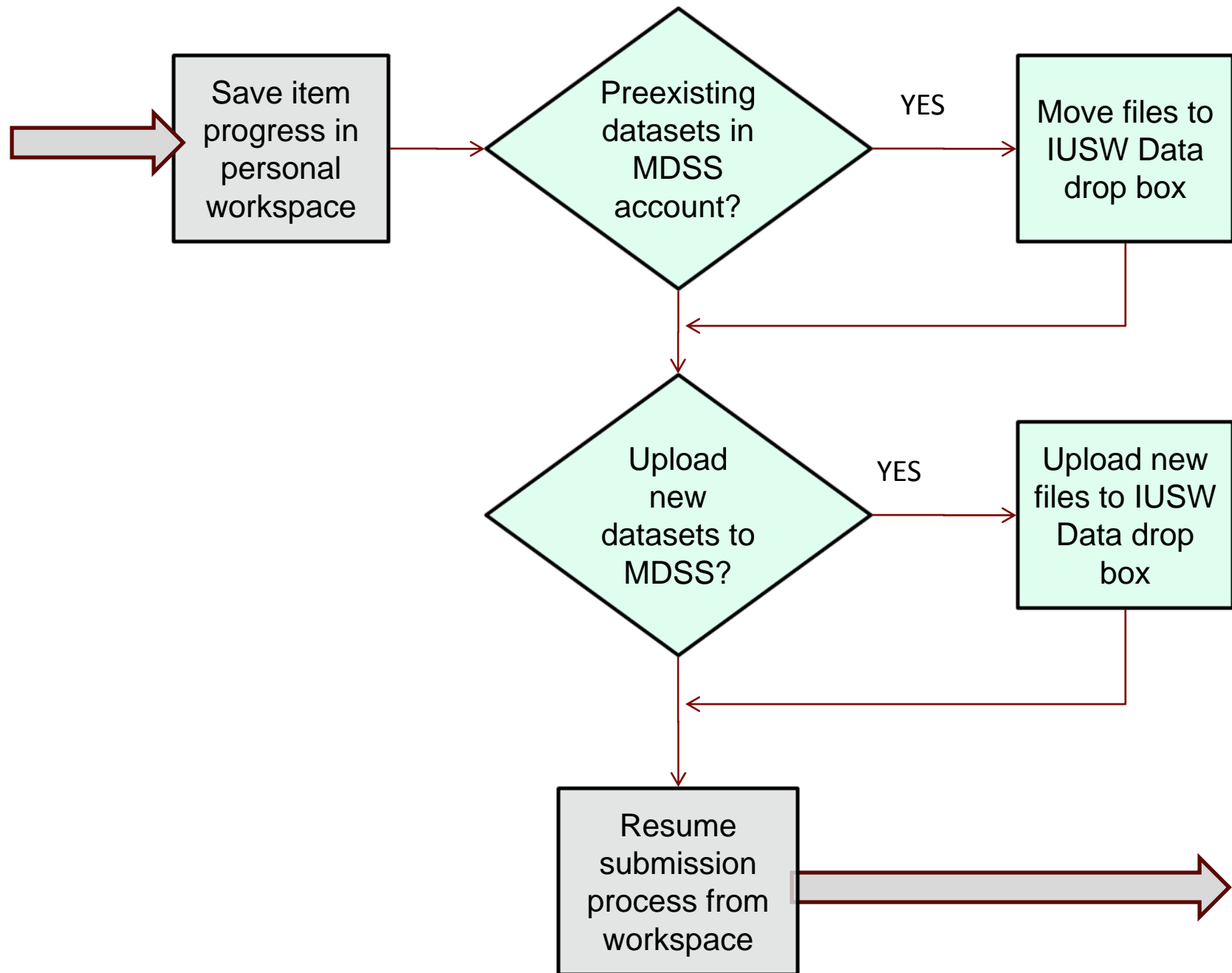
# IUScholarWorks Data in DSpace

- Facilitating the submission process for both the researcher and collection manager
  - Because some datasets are external in MDSS, this is inherently an asynchronous process for both
  - We will facilitate the process for submitters via the DSpace Configurable Submission system
  - We will facilitate the data collection manager's process via steps in the DSpace workflow system

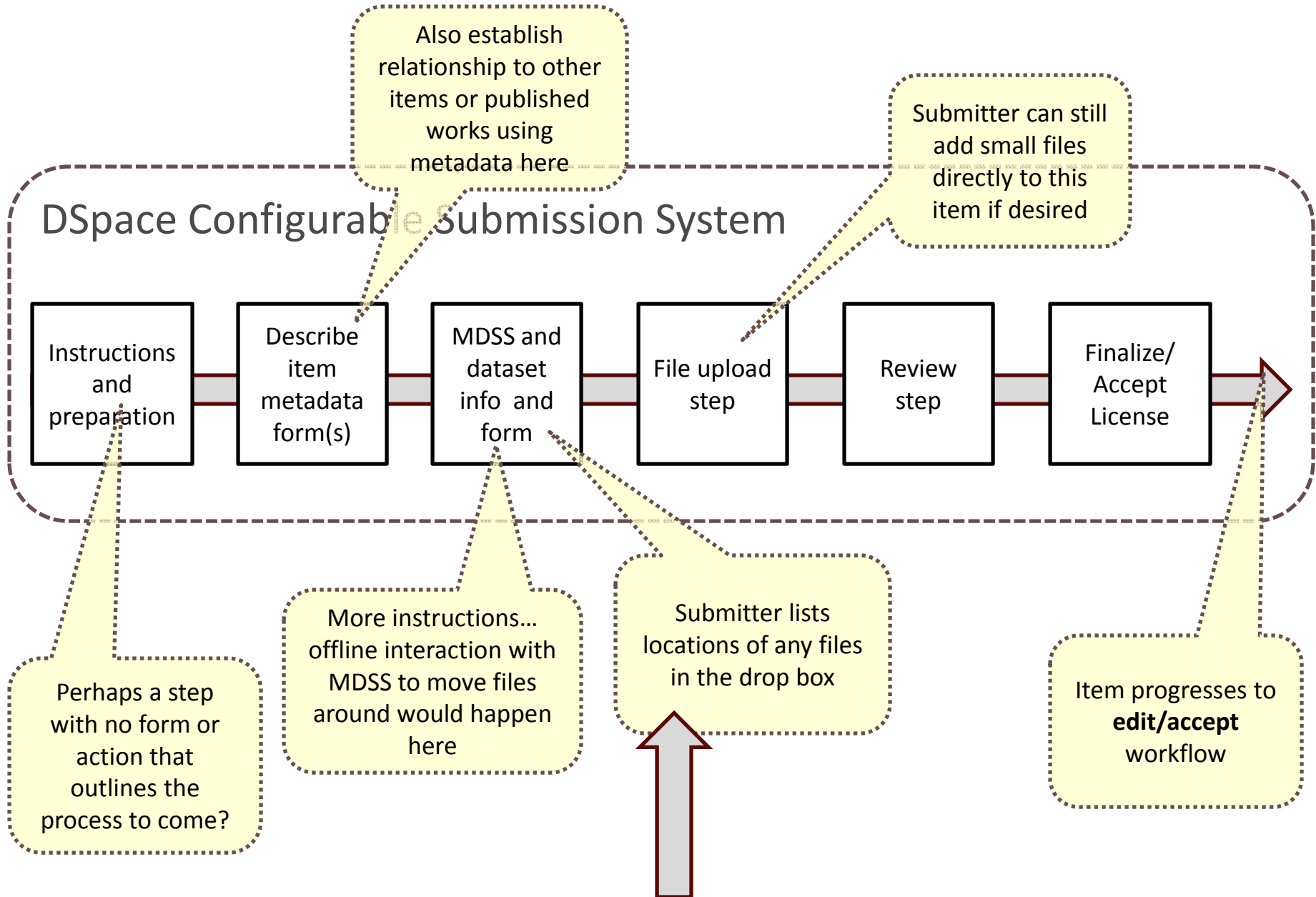
# IUScholarWorks Data: Item submission user interface



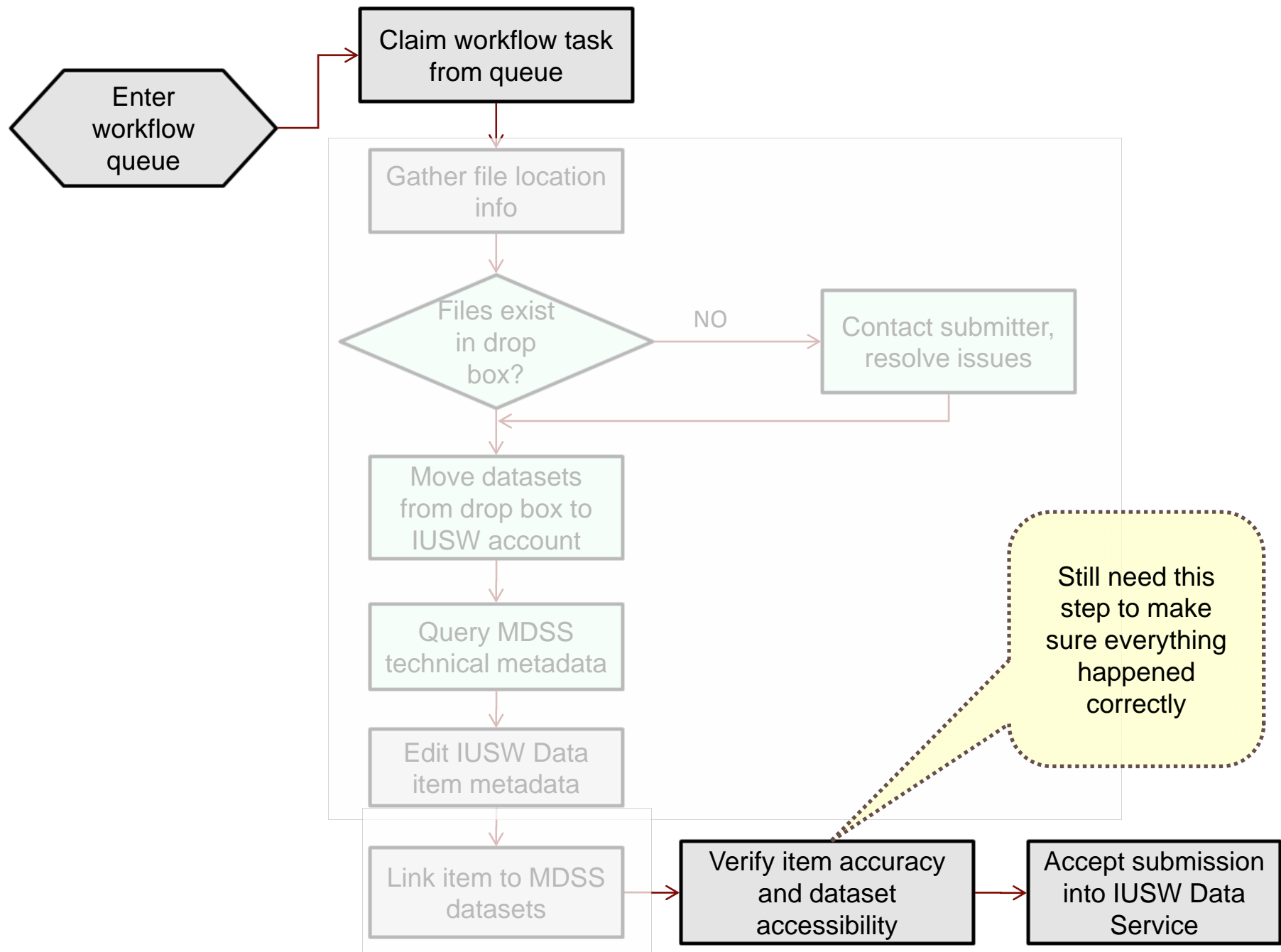
# IUScholarWorks Data: File management in IU MDSS



# IUScholarWorks Data: Item submission user interface



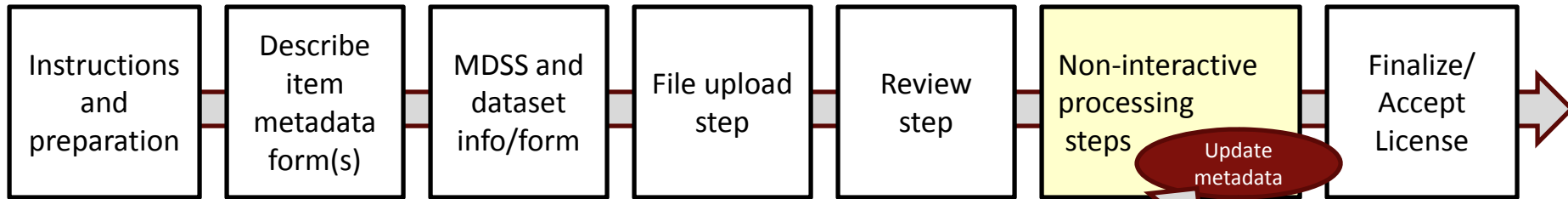
# IUScholarWorks Data: Collection Manager Workflow



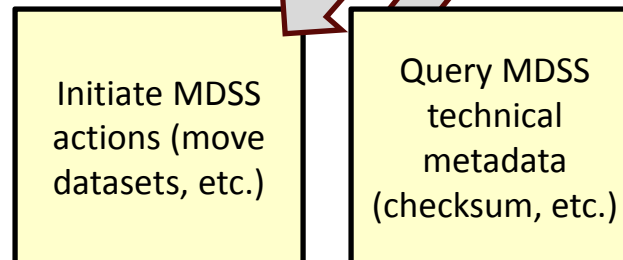
# IUScholarWorks Data: Item submission user interface

## Phase 2, automated workflow

### DSpace Configurable Submission System



### IU MDSS



# End result...

- End result is a published data item that contains:
  - Descriptive metadata
  - Links to related publications
  - Actual DSpace Bitstreams for small datasets
  - URL links to large datasets in IU MDSS
  - Technical metadata about both classes of datasets



# Beyond linking via URL...

- Storage abstraction layers to get to IU MDSS
  - DSpace support for Storage Resource Broker (SRB)
  - Akubra, a low-level storage API from Topaz and Fedora Commons
- Direct mounting of MDSS directories on the DSpace server
  - Configure a separate DSpace asset store using a network mounted filesystem from MDSS

# Beyond linking via URL...

- These solutions would all imply the same thing: configuring additional DSpace asset stores and performing *item registration*
  - We don't want to use one of those methods for the default asset store and upload very large files through the DSpace web interface

# Beyond linking via URL...

- But... item registration of existing files is a batch oriented command-line process
  - assumes ready to go packages with descriptive metadata, just like importing items

# Beyond linking via URL...

- We lose the convenience of the submission interface to facilitate the service
- The ideal solution would be to connect to IU MDSS as an alternative asset store and be able to register files **to** items through the submission interface, versus just being able to register files **as** new items

---

# Questions, opinions, or comments?